How Person Records are Maintained Within the DataHub

Collapsing Duplicate Person Records

Terms

- retained person - oldest person record determined to be a match for another person record
- retired person - most person record determined to be a match for another person record
- indicies - refers to the set of tables used in the edh for quick lookups of person ids based on different identifying factors
- activities - records provided by hospitals, call centers, Experian, etc..
- encounter - any activity provided by a hospital about a patient visit

Criteria and Steps

For a full description of the matching criteria and the steps taken after a collapse happens, see the Person Collapse documentation

Cassandra

For each table, what happens to the rows containing the information for that person?

- person_master (retired person) - retired person record deleted
- person_master (retained person) - person attributes are updated using activities from retired person
    - TODO: Logic for this is buried deep in the code and still needs to be captured and documented
- person_master_changes (retired person) - retired person record deleted
- person_master_changes (retained person) - updated during change capture process
- person_activity - the person_id on all activities for the retired person are updated to the retained person's id
- person identity indices (used for blocking keys during person identification process) - regenerated for customer
- source identity indices (used for quick lookups on source record id during identification process) - updated with new person id
- person_collapse_history - updated with the collapsed persons: customer_id, retired_person_id, retained_person_id, retired_at, blocking_key_collapsed_on

Is any functionality impaired or any data lost once the collapse occurs?

- Any reference to the retired person id in the primary schema tables (person_master and person_activity) is now gone throughout the Cassandra database and all cached copies in Redshift and Elasticsearch
- Links or references between retained and retired persons can only be made through the person_collapse_history table
- All data saved for the retired person except the fields tracked in collapse_history, or the data in snapshots and reports is lost
- Currently no list tables are updated with any information about the collaspe. Therefore lists will lose their reference to the current retained person in person_master and joins will have to be done through person_collapse_history

Redshift

For each table affected, what happens to the rows containing the information for that person?

- person_master(retired person) - deleted
- person_master(retained person) - update with new changes from Cassandra. date_modified field is set to current date time
- person_master_changes - new updates appended
- person_activity - updated with retained person id
- person identity indices (used for blocking keys during person identification process) - NA
- source identity indices (used for quick lookups on source record id during identification process) - NA
- person_collapse_history - updated with the retired persons: customer_id, retired_person_id, retained_person_id, retired_at, blocking_key_collapsed_on

Is any functionality impaired or information lost once the collapse occurs in a subsequent refresh of Redshift?

- Once the collapse occurs, the only reference to the retired person id is in the person_collapse_history table any references in reporting are lost.
- However, any copies of a retired person that were made will still remain in any existing tables such as in the case of reporting tables or list data snapshots
- Links between retained and retired persons can only be made through the person_collapse_history table
- All data saved for the retired person except the fields tracked in collapse_history, or the data in snapshots and reports is lost
- Currently no list tables are updated with any information about the collaspe. Therefore lists will lose their reference to the current retained person in person_master and joins will have to be done through person_collapse_history

Elasticsearch

For each index affected, what happens to the rows containing the information for that person?

- When generating a new index only the retained person and activity/encounter information is loaded. No reference to the original person is provided.
- Because indexes are completely replaced on each refresh, the retired person information will be removed after the next refresh
- List snapshots should still contain the original retired person information

Is any functionality impaired or information lost once the collapse occurs in a subsequent refresh of Elasticsearch?

- There is currently no mechanism to link the retired person to the retained person in the primary customer index within Elasticsearch

De-linking A Person from a Customer